# Delegation is not abdication

## An agent is governed by its gate, its contract, and its mandate

The previous instalment located an agent's security boundary in its admission gate: the mechanism that authorises or refuses its requests before execution. That was the technical question, namely what is allowed to run between two components. This instalment closes the series by shifting the focus to another question. Once the gate is in place, who answers for what the agent actually does in the world?

The intuitive answer often looks for a solution in a better model, in more attentive human supervision, or in more complete logging. Each of these answers addresses part of the problem. None resolves it entirely.

The thesis defended here is narrower. Delegating the execution of a task to an agent never transfers responsibility for its effects. But that responsibility cannot be guaranteed by the mere presence of an identified human either. An agent acting in the name of an organisation is governable only when three layers are explicitly defined: a gate that determines what it can do, a contract that determines who answers for it, and a mandate that determines what it is legitimate to delegate to it.

This thesis concerns agents that produce effects on the world: drafting deliverables, transmitting information, triggering actions, making operational decisions, interacting with third parties. It does not concern a passive assistant with no capacity to act.

The question is no longer theoretical. The EU AI Act, in its Article 14, requires effective human oversight of high-risk systems, including the ability to intervene in the system or to interrupt it.

Singapore published its Model AI Governance Framework for Agentic AI on 22 January 2026, the first national framework devoted to agentic systems. Its structure deserves attention. It explicitly distinguishes effective human accountability from technical controls, and it makes the upfront bounding of risk one of its four dimensions.

The thesis defended here therefore does not invent its distinction. It rediscovers it, independently, in the first regulatory instrument to have confronted the problem.

At the same time, case law has begun to rule. In Moffatt v. Air Canada (2024 BCCRT 149), the British Columbia Civil Resolution Tribunal rejected the company's argument that its conversational agent was a separate entity answering for its own acts, and held the organisation responsible for the information delivered. A conversational agent is not an agentic system in the sense retained here, and an administrative ruling is not a general

precedent. But the principle laid down is exactly the one defended below: responsibility is not delegated to the technical identity that executes.

## Why security and logging are not enough

Two reflexes dominate today's discussions.

- The first consists in reinforcing technical controls. This approach remains indispensable. An admission gate reduces the attack surface, limits privileges, and blocks out-of-scope behaviour. But it answers only one question: what can the system do?
- The second consists in reinforcing traceability. Every tool call, every write, every decision is recorded in an audit trail allowing a complete reconstruction of events. This approach is equally necessary. But it answers another question: what happened?

Neither provides a complete answer to the question of responsibility.

The reason is simple:

- Authorising is not imputing,
- Tracing is not answering for.

An admission gate decides which actions are permitted. A log records which actions were carried out. Neither necessarily designates who was supposed to control the action, who held the authority to delegate it, or who bears the consequences of failure.

This distinction becomes critical when several actors intervene simultaneously: the model provider, the IT team, the business owner, the end user, the organisation's leadership. In such configurations, responsibility does not disappear; on the contrary, it tends to dilute.

The relevant question is therefore not only "who triggered the action?" but also "who held the authority to delegate it?" and "who remains responsible when delegation fails?"

## From owner to mandatary

A first answer consists in requiring that a human owner be identified for each agent.

This requirement is useful but insufficient.

In simple systems, a single owner can indeed be designated. In complex systems, that representation quickly becomes artificial. Major critical infrastructures, from aviation to civil nuclear power, rarely rest on a single responsible party. They organise themselves

into distributed chains of responsibility, escalation mechanisms, and clearly defined roles.

The objective is therefore not to fictitiously assign all responsibilities to a single person. The objective is to make the structure of delegation explicit.

This distinction leads to introducing a more robust notion: that of mandate.

An agent is not merely a tool. It acts as an artificial mandatary. It receives a limited authorisation to act in the name of a person or organisation within a defined perimeter.

The governance question then becomes less "who uses the agent?" than "who granted it the power to act, and within what limits?"

This logic already exists in human organisations. An employee, a lawyer, a sales representative, or a corporate officer can act on another's behalf without thereby becoming the ultimate holders of responsibility. Agentic systems reintroduce this structure into the software world.

The analogy has a limit, however, that must be named before it is raised against us. The employee, the lawyer, the corporate officer are themselves subjects of law. They answer, at their level, for their acts. The artificial agent is the subject of nothing. It has no patrimony, no legally recognised will, no capacity to bear a sanction. This asymmetry does not weaken the thesis; it founds it. Precisely because the agent can assume nothing, responsibility never stops at it. It flows back, by construction, to whoever mandated it.

## The agent-human contract and its three clauses

The agent-human contract is the second of the three layers. It links the technical action to organisational responsibility.

It in turn breaks down into three minimal clauses, not to be confused with the three layers that encompass them.

1. The first is information and transparency. Anyone concerned must be able to know when an agent intervenes in a process producing significant effects. In some contexts this requirement will take the form of explicit consent; in others, the form of a duty to inform or a right to object. The form may vary. The principle remains.

2. The second is imputation. Each action must be linkable to an explicit chain of responsibility. The technical identity of the agent is never the terminal point of that chain. It must point back to the persons or functions that effectively hold the authority of delegation and the residual responsibility.

3. The third is control. A governable system must allow interruption, manual takeover, or limitation of its action. This capacity must not, however, be reduced to

a simple stop button. In many domains, the speed of execution makes human interruption impossible. The real objective then becomes control of the blast radius: containment, capping of effects, reversibility of operations, and limitation of privileges.

The contract therefore replaces neither the gate nor the mandate. It establishes the link between them.

## Why governance shifts toward the mandate

As agents gain autonomy, human validation of each action becomes economically impracticable.

The most interesting systems are precisely those that reduce the frequency of human intervention. This is why governance cannot rest exclusively on approval points.

The centre of gravity then shifts.

The question is no longer "does a human validate each action?"

The question becomes "which actions did one authorise the agent to undertake before they even occur?"

In other words, the governance of agents is primarily a governance of ex ante delegation.

The mandate becomes the central object. It defines the perimeter of action, the limits of autonomy, the escalation thresholds, and the conditions of suspension. It transforms a series of point-in-time authorisations into a stable framework of responsibility.

## Two fields of application

PREDICARE (a set of agents dedicated to addressing medical disengagement in the context of metabolic syndrome), a clinical decision-support agent, illustrates a case where governance rests simultaneously on the three layers.

- The gate refuses requests outside the clinical perimeter.
- The contract guarantees the practitioner's information, the imputation of recommendations, and the possibility of setting aside a proposal.
- The mandate finally defines what the agent is authorised to recommend without further validation.

OCTOPUS, a multi-model orchestration system, represents a more demanding case. Several agents contribute to the production of a single output. Technical traceability allows the chain to be reconstructed. But governance becomes effective only when the mandate clearly identifies the entity responsible for the composite output, regardless of

the number of intermediate agents involved. These examples illustrate the thesis without claiming to exhaust it.

## Limits

This position carries several limits.

The contract does not replace technical controls. A perfectly attributed responsibility does not fix a dangerous system.

The mandate does not eliminate organisational costs. Any explicit delegation introduces trade-offs, delays, and supervision mechanisms.

Direct human control quickly becomes impracticable when execution times are measured in milliseconds. In such cases, reversibility and containment replace interruption.

Finally, responsibility remains distributed in complex systems. Governance does not eliminate this distribution; it makes it visible.

## Implication for the executive committee

The strategic error would be to expect the next model, or the next logging system, to solve a problem that is first of all one of governance.

An audit log records who signed. It has never said who should have.

The fundamental question bears neither on what an agent can do, nor on what it has done. The gate and the logs already answer those. It bears on who holds the authority to delegate the action, within what limits, and under what residual responsibility.

From the moment an agent acts in the name of an organisation, it becomes a new subject of governance. The decision then ceases to be merely technical. It touches the delegation of authority, legal responsibility, and corporate governance.

An agent is governable only when its architecture explicitly links three things: what it can do, what it is permitted to do, and who answers for it. Execution can be delegated. Responsibility never can.